图书情报工作
第62卷 第12期 2018年6月

面向摘要结构功能划分的模型性能比较研究*

王东波[1] 陆昊翔[1] 周鑫[2] 朱丹浩[3]

[1] 南京农业大学信息科学技术学院 南京 210095 [2] 南京大学信息管理学院 南京 210093

[3] 南京大学计算机科学与技术系 南京 210093

摘要：［目的/意义］摘要作为学术论文中能够简明扼要地说明研究目的、研究方法和最终结论的陈述部分，具有较高的探究价值和意义。［方法/过程］选取长短期记忆网络（Long Short-Term Memory）、支持向量机（Support Vector Machine）、LSTM-CRF 和 CNN-CRF 4 种模型，对 3 672 篇情报学领域的期刊论文进行摘要划分识别研究。［结果/结论］长短期记忆网络模型识别 F 值最高为 69.15%，LSTM-CRF 神经网络模型最高 F 值为 88.76%，RNN-CRF 模型最高 F 值达到 89.10%，支持向量机分类器分类宏观 F 值最高为 72.04%。该实验结果对图书情报领域的学术论文结构功能划分实验模型选取有较高的参考价值。

关键词：结构功能划分 条件随机场 长短期记忆网络 卷积神经网络 支持向量机

分类号：G255.1

DOI：10.13266/j.issn.0252-3116.2018.12.011

1 引言

摘要是对学术论文不加注释和评论的简短陈述，基本要素包括研究目的、方法、结果和结论，是具有独立性和完整性的短文[1]。摘要作为学术研究中的重要组成部分，能够为研究人员提供完整的学术文献主要信息，在无法获取学术论文全文信息及全文数据处理困难的情况下，它成为最具有研究价值的数据来源之一。在机器学习技术迅速发展的前提下，如何从摘要中挖掘出相应的知识成为面向学术文本进行深度知识挖掘的重要研究内容之一，而根据摘要已有的结构划分标记构建摘要结构功能自动划分模型是进行上述深度知识挖掘的基础。在上述研究背景下，基于不同的机器学习模型，笔者构建了面向摘要的不同种类的结构功能划分模型，并对不同模型的性能进行了对比和分析。不仅为验证不同机器学习模型在摘要结构功能自动划分上的性能状况提供了第一手的资料，而且为面向摘要进行结构功能划分确定了最优的模型，从而为进行全文本的结构功能划分提供了相应的模型借鉴。

目前，已经有一些学者从文本结构的划分和机器学习[2]的角度对相关的研究进行调研。陆伟等[3]采用条件随机场模型，基于章节的标题对学术文本的结构功能进行识别实验，取得了较好的实验结果。这一研究把条件随机场模型有机地融入到学术文本的篇章结构自动识别当中，充分利用了标题中的特征词，在研究方法上具有较强的可借鉴性。黄永等[4,5]通过构建支持向量机分类器分别基于章节内容和段落内容对学术论文的结构功能进行识别，达到了较高的准确率。虽然该研究使用了通用的分类模型，但从领域应用的角度看这一研究具有较强的创新性。崔建明等[6]通过引入“粒子群算法”，对 SVM 算法进行改进，在文本分类实验中提高了原始 SVM 分类器的性能。这一研究在 SVM 分类模型中融入了算法特征，为该模型提供了确切的特征知识，提高了整个模型的性能。在融合两个机器学习模型的基础上，程健一等[7]通过构建 SVM 和 CRF 双层分类器，实验 F 值达到 91.1%。这一研究充分利用了线性和非线性两个模型共有的优势，具有方法上的创新性。胡新辰[8]提出一个基于 LSTM（长期短期记忆）的深度学习模型来解决语义结构关系分类问题，并在标准评测集合上取得的成绩达到了当时的最好水平。首次把 LSTM 应用在语义结构关系分类的

* 本文系国家社会科学基金重大项目“情报学学科建设与情报工作未来发展路径研究”（项目编号：17ZDA291）研究成果之一。

作者简介：王东波（ORCID：0000-0002-9894-9550），副教授，硕士生导师；陆昊翔（ORCID：0000-0002-6855-6393），本科生；周鑫（ORCID：0000-0001-7756-2253），博士研究生；朱丹浩（ORCID：0000-0003-0477-8517），助理馆员，博士研究生。

收稿日期：2017-12-16 修回日期：2018-04-01 本文起止页码：84-90 本文责任编辑：徐健

研究上,对于如何更好地发挥 LSTM 在语义分类上的整体性能也进行了细致的探究。任智慧等[9]在序列标注实验中,提出基于 LSTM 网络模型的改进方法,采用六词位字符标注集并加入预先训练的字嵌入向量(字符嵌入)进行中文分词,证明基于 LSTM 网络模型的方法比当前传统机器学习方法具有更好的性能。从整个研究内容看,六词位字符集的确定对于其他的研究如何确定面向深度学习的字符集具有较强的借鉴性。针对具体的分词任务,张子睿和刘云清[10]提出了一种基于长短期记忆神经网络改进的双向长短期记忆条件随机场(BI-LSTM-CRF)模型。这一研究的创新之处在于基于深度学习模型的性能优势,充分挖掘了由字构词过程中所使用的字的左右特征,进而确保了所构建模型的整体性能。J. P. C. Chiu 和 E. Nichols 基于 CNN-CRF 模型通过计算字符级别的特征实现领域内实体识别,取得了很好的效果[11]。这一研究在识别实体的过程中所使用的字符及其周围的特征对于本文的研究具有直接的借鉴意义和价值。

基于上述已有的相应研究,在学术论文摘要部分功能结构划分越来越规范的背景下,笔者选取来自"中国知网"图书情报学领域核心期刊中具有明确摘要功能划分的 3 672 篇学术论文构建语料库,基于机器学习算法和深度学习理论搭建了 4 种机器自动分类模型,分别对 3 672 篇学术论文摘要部分实现功能结构的自动划分,通过对 4 种模型分类识别效果的比较展示了相关领域结构功能划分识别中机器学习模型的优劣性。笔者采用两种结构划分识别理念进行实验:①从"序列标注"的角度进行识别实验,在这种实验思想下进行了"长短期记忆网络标注实验""添加 CRF 层的长短期记忆网络标注实验"和"外接 CRF 层的卷积神经网络标注实验",这 3 个实验均选择独立的字作为最小的处理单元;②从"整体分类"的角度进行识别,在该理念下进行了"构建支持向量机分类器实验",将每一独立功能结构作为最小处理单元。具体实验流程见图 1。

2 模型介绍

2.1 长短期记忆网络

长短期记忆网络(Long-Short Term Memory)是一种经过特殊设计的 RNN(Recurrent Neural Network)模型,能够学习长期的依赖关系[12]。从摘要自动分类的这一具体任务来看,长短期记忆网络不仅适应于摘要句子过长的这一特征,而且在一定程度上可以保持内部

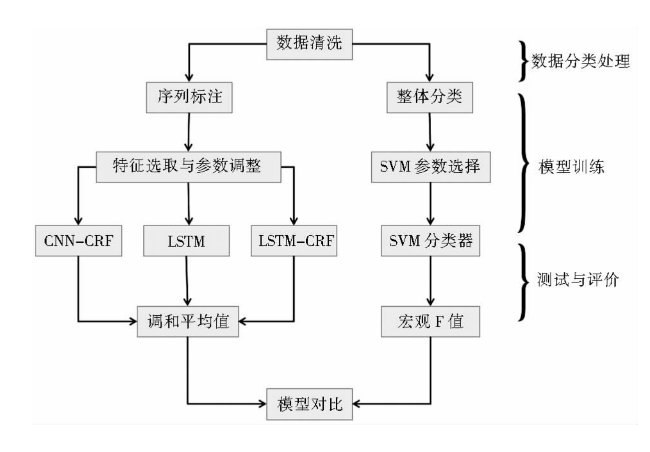


图 1 实验流程

梯度不受不利特征变化的干扰。

LSTM 单元由记忆单元(memory cell)和多个调节门(gate)组成,input gate(输入门)、output gate(输出门)和 forget gate(遗忘门)协同控制信息的输入、输出和丢弃,输入门决定哪些信息被神经元接受,遗忘门决定哪些历史信息被保留和删除,输出门决定哪些信息被输出到下一记忆单元中。"馆"字向量进入神经单元中,输入门允许部分向量信息进入神经元,同时遗忘门删去该字与"基""于"字向量的联系,输出门将"馆"与"图""书"等字向量的关联度,及"馆"字出现位置、是否构成词语等信息传入下一神经单元,确保"馆"字与历史字符关联性信息和其他字符级特征得到保存。

2.2 LSTM-CRF 模型

条件随机场模型[13]的不足之处在于,为了得到更好的识别效果,需要人为地寻找和添加数据特征,这对研究人员的数据敏感性有着较高要求,而且大量隐藏特征无法被识别,使得模型的性能不能得到最优化的体现。LSTM 神经网络的优势在于挖掘深层的隐藏语义关联性,并以向量的形式表现出来。在实际的序列标注任务中,由于神经网络结构对数据的依赖性很强,数据量的大小和质量都会严重影响模型训练的效果。

LSTM-CRF 模型可以很好地解决这一问题,数据经过 LSTM 网络的处理,最终输出的向量即可以看成是输入数据的一种表示形式,LSTM 挖掘深层特征信息后导入条件随机场模型中,使模型特征的质量得到较大提升。新模型综合利用两种模型的优势,LSTM 层解决了提取序列特征的问题,CRF 层有效利用了句子级别的标记信息。从理论和方法论上分析,这一组合模型具有特定的优势。图 2 示例为"目的"类别部分文本序列示例:

该模型输入文本观测序列,LSTM 层通过单个记忆

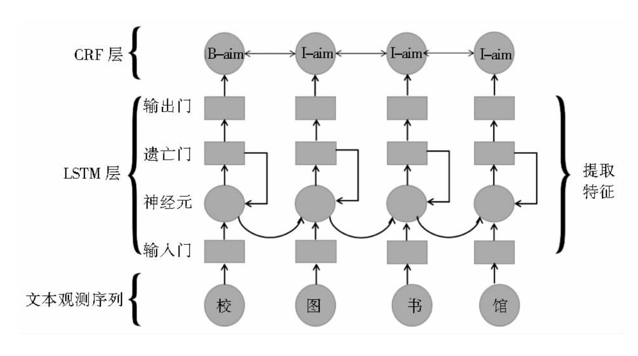


图2 LSTM-CRF模型示例

单元中3种gate的协作,有效利用上下文语义联系提取文本中的深度特征信息,以向量的形式输入CRF层,经过条件概率的计算,输出文本标签。"校""图""书"和"馆"观测序列以单字符格式进入LSTM网络,独立神经单元处理提取字符级特征及该字符与历史字符的关联信息,全部数据信息以向量形式被条件随机场模型处理,模拟手动建立特征模板和数据特征输入,计算出字符间条件概率矩阵,判断"校"字为功能块首字,"图""书"和"馆"作为功能块中间字符,最终得到"B-aim I-aim I-aim I-aim"序列。

2.3 CNN-CRF模型

卷积神经网络(Convolutional Neural Network, CNN)是一种前馈式神经网络[14],设计之初是对大型图片进行处理,其独特的卷积结构和信息反馈机制能确保其在文本分类领域也得到了广泛应用。由于CNN在一定程度上不仅能够充分进行显示特征抽取,而且可以隐式地从训练数据中进行特征的自我抽取,所以其构建的分类模型在性能上具有突出的优势。与LSTM-CRF模型相同,CNN网络的作用在于通过卷积核的卷积提取深层字符级特征,协助CRF模型对文本进行分类标识。具体模型工作情况如图3所示:

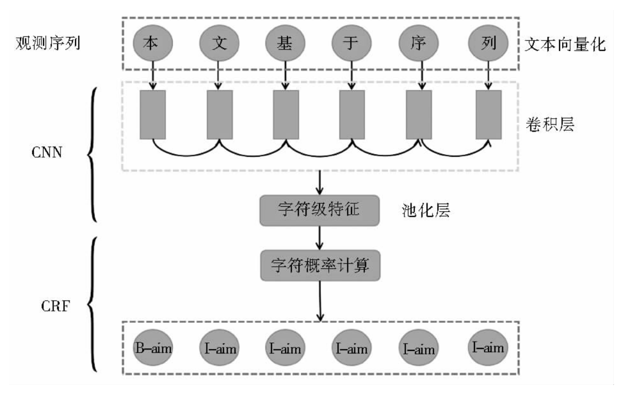


图3 CNN-CRF模型示例

观测序列经过数据向量化过程,以单个字符为单位生成字表矩阵,CNN卷积核以单个字符为中心,对该字符周围字符进行卷积,提取该局部特征和该特征与其他特征之间的位置信息,循环式的卷积操作提取全部局部信息,池化层整合全部局部信息,经过计算获得字符级特征信息;字符级信息以向量的形式进入CRF层,通过字符特征概率计算,得到最终的序列标注。如图3所示,若卷积核大小为2,当卷积核以"基"为卷积中心,左右距离小于等于2的字符将被纳入卷积计算中,在提取特征信息之后,卷积核转移到"于"字,重复以上操作。循环进行的卷积过程将得到提取全部局部特征,池化层整合所有局部特征,得到全文特征信息,以向量形式被条件随机场处理,得到"B-aim I-aim I-aim I-aim I-aim"序列。

2.4 支持向量机

支持向量机(SVM - Support Vector Machine)是机器学习中支持向量计算的分类器,也一种优秀的有监督学习算法,其核心内容是在1992到1995年间提出的[15],目前仍处在不断发展阶段。

SVM是从线性可分情况下的最优分类面发展而来的[16],在训练数据中每个数据都有n个的属性和一个二类类别标志,可以认为这些数据在一个n维空间里。我们的目标是找到一个n-1维的超平面(hyperplane),这个超平面可以将数据分成两部分,每部分数据都属于同一个类别。类别中距离分类面最近的平面上的数据为支持向量,当类别之间支持向量所在平面距离最远时,平行平面的中间平面即为最优分类平面。

3 数据处理

在2017年11月3号到10号之间,笔者从CNKI数据库中获取了2014-2017年间《图书情报工作》《情报杂志》《情报探索》《数据分析与知识发现》(原《现代图书情报技术》)《情报科学》《情报理论与实践》《现代情报》和《农业图书情报学刊》等期刊上含有结构功能标记的摘要,共计3 672篇。首先对语料进行一致性清洗工作,原始语料中对于摘要功能块的标注符号分为"【】"和"[ ]"两种,将所有标注符号统一为"【 】";结构单元的标识词内部对于空格及其他分隔符号的使用不统一,清洗过程中将分隔符统一为单空格。

笔者从两种不同的角度对功能结构进行自动识别,根据对应角度对数据做不同处理。

在"序列标注"标注实验下,以单独的字为处理单元,"目的""方法""结果"和"局限"功能结构分别用"aim""med""con"和"lit"标识,独立功能结构作为整

王东波，陆昊翔，周鑫，等.面向摘要结构功能划分的模型性能比较研究[J].图书情报工作,2018,62(12):84-90.

体添加入实验数据,同时每一种功能结构的段落首字用“B-”标识,段落中字用“I-”标识,段落尾字用“E-”标识,一共有12种标签,具体如图4所示:

多 B-aim 媒 I-aim 体 I-aim 时 I-aim 代 I-aim 用 I-aim 户 I-aim 的 I-aim 信 I-aim 息 I-aim 行 I-aim 为 I-aim 。 E-aim

本 B-med 文 I-med 提 I-med 出 I-med 一 I-med 种 I-med 新 I-med 的 I-med 模 I-med 型 I-med ，E-med

聚 B-con 类 I-con 准 I-con 确 I-con 度 I-con 要 I-con 优 I-con 于 I-con 传 I-con 统 I-con 模 I-聚 I-con 方 I-con 法 I-con 。 E-con

利 B-lit 用 I-lit 本 I-lit 体 I-lit 实 I-lit 现 I-lit 语 I-lit 义 I-lit 时 I-lit 未 I-lit 将 I-lit 本 I-lit 体 I-lit 进 I-lit 行 I-lit 集 I-lit 成 I-lit 。 E-lit

图4　序列标注示例

在独立LSTM序列标注实验、LSTM-CRF序列标注实验和CNN-CRF序列标注实验中,均在模型中添加了语料库整体字向量信息,该字向量是由Word2vec算法[17]构建的100维向量。

在“整体识别”实验下,将具有独立功能结构的段落作为整体构建文本向量。笔者通过字向量搭建文本向量,常用的构建文本向量的算法有Word2vec算法和TF-IDF(Term Frequency-Inverse Document Frequency)算法[18],在计算字向量[19]时选择TF-IDF算法作为核心算法,将语料根据功能划分为目的、方法、结果和局限4个部分,计算语料中所有单字(包括汉字、数字和符号)的TF-IDF权重(单个字的权重 = 4种类别中该字的TF-IDF权重 * 对应类别的数据数量之和),选择权重最高的150词作为基准词,逐一计算结构单元中150个基准词的出现频次,构建出150维的字向量,并搭建出语料文本向量。表1展示了部分综合TF-IDF权重较高的基准单字。

表1　部分基准单字示例

序号	基准单字	综合TF-IDF权重
1	用	43.93
2	对	42.48
3	研	41.16
4	分	41.15
5	和	39.50
6	数	34.97
7	究	34.74
8	学	33.35
9	进	33.09
10	为	31.74
11	行	30.37
12	信	30.19
13	在	28.83
14	据	28.74
15	方	27.99

为科学详细展示模型分类性能,笔者在构建训练文本和测试文本时进行了三组对比实验,训练语料与测试语料数据量比例分别为7:3、8:2和9:1,较为全面地展示了模型在不同数据量环境下的分类结果。

4　实验结果及评价

对于序列标注实验的评价标准有三个指标,分别为准确率(Precision)、召回率(Recall)、F值(F-measure),整体识别实验评价标准增加了宏平均值(Macro-average)作为所有类别整体识别评价指标,具体计算公式如下:

$$准确率\ P = \frac{A}{A+B} \times 100\% \qquad 公式(1)$$

$$召回率\ R = \frac{A}{A+C} \times 100\% \qquad 公式(2)$$

$$调和平均值\ F = \frac{2 \times P \times R}{P+R} \times 100\% \qquad 公式(3)$$

$$宏平均值\ F = \frac{\sum_{i=1}^{n} F_i}{n} \qquad 公式(4)$$

其中,A表示功能块识别正确的个数,判断标准为段落的首字识别正确且段落尾字识别正确,则该功能块识别正确。B表示错误识别功能块的个数,C表示未识别出的功能块的个数,F_i为各类别独立的调和平均值。为确保实验数据的准确性和理论的科学性,笔者选择进行10折交叉验证(10-fold cross-validation)实验[20]。

4.1　长短期记忆网络序列标注实验

本实验使用Python程序语言,在搭载4GB显存“NVIDIA”Quadro K1200型GPU和“英特尔酷睿i5-4590”四核处理器的Linux操作系统下,基于Tensorflow框架[21]搭建了深度神经网络,模型中可修改是否在神经网络中添加CRF网络层。本实验在未添加CRF层条件下基于Tensorflow框架进行LSTM序列识别,选择“adam”作为模型优化器,文本向量与LSTM隐藏单元数量均为100。具体十折交叉结果如表2所示:

表2　长短期记忆网络模型十折交叉实验

实验序号 \ 实验比例	9:1实验F值	8:2实验F值	7:3实验F值
1	68.05%	57.62%	48.84%
2	68.32%	57.67%	48.87%
3	67.67%	57.98%	49.49%
4	68.12%	58.43%	48.68%
5	68.95%	58.16%	49.11%
6	68.65%	57.78%	48.73%
7	68.86%	58.59%	49.09%
8	69.15%	57.38%	48.68%
9	68.28%	58.05%	49.49%
10	68.00%	57.94%	49.17%
均值	68.40%	57.96%	49.02%

独立 LSTM 序列识别效果的宏观 F 值在训练测试比为 9:1 情况下最高为 69.15%，均值为 68.40%，当减少训练语料数据量至70%时宏观 F 值只有49.02%。总体识别情况较差，主要影响因素为摘要功能块文本过长，不同于一般的实体名称识别，而且单字的标签极大地依赖于相邻字的标签和整体序列信息，长短期记忆网络在挖掘深层特征之后无法有效实现序列识别和标注，实验结果表明 LSTM 模型在类似结构功能划分识别实验中还有较大改进空间。

4.2　LSTM-CRF 模型序列识别实验

与独立 LSTM 实验相同，修改基于 Tensorflow 框架的神经网络参数，添加 CRF 为神经网络最后一层，将 LSTM 层挖掘的深层特征向量作为 CRF 层的输入量。具体实验结果如表 3 所示：

表 3　LSTM-CRF 模型十折交叉实验

实验序号 \ 实验比例	9:1 实验 F 值	8:2 实验 F 值	7:3 实验 F 值
1	88.13%	87.42%	86.54%
2	88.42%	87.60%	86.87%
3	87.57%	87.58%	86.31%
4	88.58%	88.15%	86.48%
5	88.78%	88.16%	87.11%
6	88.56%	87.78%	86.73%
7	88.86%	88.09%	86.22%
8	89.05%	87.38%	86.38%
9	88.17%	88.05%	86.49%
10	88.29%	87.32%	87.17%
均值	88.44%	87.75%	86.63%

从表 3 中可以看出，添加了 CRF 层的神经网络识别效果有了很大的提升，训练测试比为 9:1 条件下最佳 F 值达到了 88.44%，基本达到预计的效果。但是距离用于实际开放性测试还有一定差距，最重要的制约因素是语料的规模，神经网络对于特征的挖掘和词间关系计算得出的概率矩阵较简单，使输入至 CRF 层的向量中包含的语义信息过少，识别效果降低，当语料规模扩大，神经网络模型对于数据处理效果理论上会有较大提升。

4.3　CNN-CRF 模型序列识别实验

本实验基于 Tensorflow 框架搭建了含卷积网络层和 CRF 层的神经网络模型，CNN 层为特征筛选层，进过处理的字符串数据由 CRF 计算条件随机概率，对文本进行标注，具体实验过程见表 4。

在本实验中，CNN 网络层的作用与 LSTM-CRF 模型中 LSTM 网络的作用类似，提取深层语义特征信息

表 4　CNN-CRF 模型十折交叉实验

实验序号 \ 实验比例	9:1 实验 F 值	8:2 实验 F 值	7:3 实验 F 值
1	88.76%	86.83%	85.35%
2	88.90%	86.18%	85.87%
3	88.77%	86.67%	85.43%
4	88.68%	87.28%	86.13%
5	88.79%	86.22%	85.82%
6	88.78%	86.80%	85.55%
7	88.66%	86.70%	85.92%
8	89.10%	87.11%	85.37%
9	88.97%	87.22%	86.09%
10	88.89%	86.57%	86.17%
均值	88.83%	86.76%	85.77%

输入到 CRF 层，弥补条件随机场模型严重依赖特征的不足。在实验所用的摘要语料中，卷积神经网络特征获取性能与长短期记忆网络的几乎处于同一水平，实验最佳 F 值 89.10%，证明了 CNN 网络在处理文本功能分类实验中具有很好的性能。

4.4　支持向量机分类器实验

SVM 中最重要的两个参数为 C 和 gamma。C 是惩罚系数，即对误差的宽容度。C 越高，说明越不能容忍出现误差。C 过大或过小，泛化能力变差，惩罚系数理论上越大越有效，但是 C 过大可能引起数据过度拟合。gamma 是选择核函数作为 kernel 后，该函数自带的一个参数，隐含地决定了数据映射到新的特征空间后的分布，gamma 越大支持向量越少，gamma 值越小支持向量越多。支持向量的个数影响训练与预测的速度[22]。核函数能提高模型的 Feature 维度（低维到高维），使 SVM 具有较好的非线性拟合能力，核函数常用的有 linear、poly 和 rbf，为了选择最优参数对模型进行测试，在训练语料与测试语料数据量比例为 9:1 背景下进行了多次组合实验得到最优参数。具体实验结果如表 5 所示：

表 5　支持向量机参数选择

代价函数（C）	核函数（Kernel）	测试语料数据量	识别正确的数量	宏平均值
1.0	linear	1 130	792	70.0%
0.8	linear	1 130	810	71.7%
1.0	rbf	1 130	414	36.6%
0.8	rbf	1 130	375	33.2%
1.0	poly	1 130	767	67.9%
0.8	poly	1 130	767	67.9%

根据实验结果，选择核函数 kernel = linear，代价函数 C = 0.8。在模型性能最优参数下进行十折交叉

实验，支持向量机的评价标准为准确率，即正确识别功能块的个数与测试语料中全部功能块的数量的比值。具体实验结果如表 6 所示：

表 6　支持向量机分类结果

实验序号 \ 实验比例	9:1 宏平均值	8:2 宏平均值	7:3 宏平均值
1	71.05%	70.82%	69.40%
2	71.29%	70.72%	70.23%
3	70.98%	69.89%	69.77%
4	72.04%	70.89%	68.86%
5	71.93%	71.23%	70.33%
6	71.56%	69.78%	69.66%
7	70.86%	70.56%	69.85%
8	71.23%	70.78%	69.56%
9	71.46%	71.19%	70.44%
10	72.00%	71.11%	68.74%
均值	71.44%	70.70%	68.68%

支持向量机的总体识别结果不佳，最高识别准确率 72.04%，低于实验预期 80% 准确率。笔者从功能块相似度角度分析了所有语料分词和功能结构分布，具体过程为调用中国科学院“NLPIR 汉语分词系统”，对全部语料进行分词处理，去除停用词后分别选取“目的”“方法”“结果”和“局限”4 种类别中词频最高的 50 个词，统计分布位置非单一的词语，一共有 78 个多类别分布的词语，如“效果”“过滤”“启示”“指标”“设计”“模板”和“人群”等词，同时高频出现在多个类别中。SVM 作为典型的小样本学习方法，对学习结果起决定作用的是少数落在分类超平面两侧的支持向量，因此本质还是对于分类类别特征的提取。在语料相似度较高时，各类别浅层特征不明显，导致支持向量的数量过少，影响了支持向量机分类器的性能。上文中提及的“效果”“过滤”和“模板”等词语出现次数均在 5 万次至 8 万次之间，研究人员对于摘要部分各类别之间互相补充说明的方式使得类别区分度降低，支持向量数量减少，SVM 分类决策失误增多，准确率降低。例如，“【目的】信息网络中的作品、个人信息等利用一般需要获得许可，而明示许可基本无法实现。默示许可可以弥补网络信息利用中明示许可存在的不足，在一定程度上解决网络信息利用中的授权问题。【方法】文章对网络信息利用中的默示许可问题进行系统讨论。【结果】从法理和实践两方面阐述默示许可在网络信息利用中适用的可行性，提出默示许可的适用范围、适用条件和适用限制。”“目的”和“结果”部分识别度较低，特征提取较困难。

5　结语

笔者从自然语言处理模型选取角度出发，对比“长短期记忆模型”“LSTM-CRF 模型”“CNN-CRF 模型”和“支持向量机”4 种经典机器学习模型，实验结果表明语料库数据较少时，基于序列标注思想的“神经网络 + 条件随机场”在处理文本结构功能划分识别问题时仍具有一定优势，独立 LSTM 神经网络模型无法有效处理序列标记识别问题，而添加 CRF 隐藏层的 LSTM 神经网络模型相较于独立 LSTM 模型性能会有较大提升，囿于语料库的数据量，深层神经网络的优势未得到体现，SVM 模型在类别识别度较低的分类问题中，代替向高维空间的非线性映射的核函数表现不佳，最佳分类超平面的选取较为困难，整体模型性能也受到影响。在针对摘要这一结构功能划分的任务上，机器学习模型特别是深度学习下的神经网络模型不仅能够有效地利用摘要句子之间和字与字之间的特征，而且在多分类的任务上确实表现出来非常强的性能优势，但如何解决机器学习过程中的领域过拟合性和迁移性是本研究在后续探究中应该关注的一个问题。

在下一步的研究中，笔者将在语料选择和模型处理方面做出更多改进，选取的语料为中文摘要数据，虽具有一定的代表性，但与英文数据中语义信息和语法构成存在较大差距，笔者将在英文摘要语料和论文正文部分进行结构功能划分机器学习模型比较实验，得到更具有普遍性的模型性能对比数据。在模型处理方面主要体现为增加语料库数据，并重新训练 LSTM-CRF 神经网络模型测试模型性能；调整卷积神经网络中卷积核大小，反馈式测试 CNN-CRF 模型的性能；使用 Word2vec 模型训练文本向量，结合长短期记忆模型提取类别特征，提高支持向量机的分类准确率；在神经网络中考虑加入 SVM 层作为特征提取层，弥补序列标注实验中对于类别边界特征利用不足的情况，进一步增加模型的数量并完善模型的识别性能。

参考文献：

[1] 刘雅琴，蒋菡，苏亚志. 科技论文摘要写作中的一些问题及辨析[J]. 现代情报，2004，24(1):178.

[2] CUNDALL P A，STRACK O D L. A discrete numerical model for granular assemblies[J]. Geotechnique，1979，29(1):47－65.

[3] 陆伟，黄永，程齐凯，等. 学术文本的结构功能识别功能框架及基于章节标题的识别[J]. 情报学报，2014 (9):979－985.

[4] 黄永，陆伟，程齐凯. 学术文本的结构功能识别——基于章节内容的识别[J]. 情报学报，2016，35(3):530－538.

[5] 黄永，陆伟，程齐凯，等. 学术文本的结构功能识别——基于段

落的识别[J]. 情报学报,2016,35(5):530 – 538.

[6] 崔建明, 刘建明, 廖周宇. 基于 SVM 算法的文本分类技术研究[J]. 计算机仿真, 2013, 30(2):299 – 302.

[7] 程健一, 关毅, 何彬. 基于 SVM 和 CRF 双层分类器的英文电子病历去隐私化[J]. 智能计算机与应用, 2016, 6(6):17 – 19.

[8] 胡新辰. 基于 LSTM 的语义关系分类研究[D]. 哈尔滨: 哈尔滨工业大学, 2015.

[9] 任智慧, 徐浩煜, 封松林,等. 基于 LSTM 网络的序列标注中文分词法[J]. 计算机应用研究, 2017, 34(5):1321 – 1324.

[10] 张子睿, 刘云清. 基于 BI-LSTM-CRF 模型的中文分词法[J]. 长春理工大学学报(自然科学版), 2017, 40(4):87 – 92.

[11] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016,10(4):357 – 370.

[12] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735 – 1780.

[13] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Eighteenth international conference on machine learning. San Francisco: Morgan Kaufmann Publishers Inc. ,2001: 282 – 289.

[14] KARPATHY A, JOHNSON J, FEI F L. Visualizing and understanding recurrent networks[EB/OL]. [2017 – 10 – 12]. https://arxiv. org/abs/1506.02078v2 2015.

[15] CORTES C, VAPNIK V. Support-Vector networks[J]. Machine learning, 1995, 20(3):273 – 297.

[16] 刘华煜. 基于支持向量机的机器学习研究[D]. 大庆: 大庆石油学院, 2005.

[17] 唐明, 朱磊, 邹显春. 基于 Word2Vec 的一种文档向量表示[J]. 计算机科学, 2016, 43(6):214 – 217.

[18] 黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报, 2011, 34(5):856 – 864.

[19] 廖健, 王素格, 李德玉,等. 基于增强字向量的微博观点句情感极性分类方法[J]. 郑州大学学报(理学版), 2017, 49(1): 39 – 44.

[20] 牛晓太. 基于 KNN 算法和 10 折交叉验证法的支持向量选取算法[J]. 华中师范大学学报:自然科学版, 2014, 48(3):335 – 338.

[21] ABADI M, AGARWAL A, BARHAM P, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems[EB/OL]. [2017 – 11 – 12]. https://arXiv:1603.04467.

[22] SU C T, YANG C H. Feature selection for the SVM: an application to hypertension diagnosis[J]. Expert systems with applications, 2008, 34(1):754 – 763.

作者贡献说明:

王东波:总体设计、数据结果分析和撰写;

陆昊翔:模型训练和论文撰写与修改;

周鑫:模型调参;

朱丹浩:深度学习模型搭建。

A Comparative Study of Model Performances Facing Abstract Structure Function

Wang Dongbo[1] Lu Haoxiang[1] Zhou Xin[2] Zhu Danhao[3]

[1] College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095

[2] Department of Information Management, Nanjing University, Nanjing 210093

[3] Department of Computer Science and Technology, Nanjing University, Nanjing 210093

Abstract: [**Purpose/significance**] Abstract can explain concisely the research purposes, research methods and the final part of the statement, which is of high exploration value and significance. [**Method/process**] In this paper, four short-term memory networks (long short-term memory, support vector machine, LSTM-CRF and CNN-CRF) were selected to summarize the journal articles of 3672 CNKI databases. [**Result/conclusion**] The long-term memory network model identifies the highest F value of 69.15%, the maximum F value of LSTM-CRF neural network model is 88.76%, and the highest F value of RNN-CRF model is 89.10%. The highest support vector machine classifier classification macro F value is 72.04%. The experimental results have a high reference value for the selection of the experimental model of the functional structure of academic dissertation in the field of library and information science.

Keywords: structure function division condition random field long-term memory network convolutional neural network support vector machine